CS 123

EMPIRICAL EVIDENCE FOR A PROPOSED DISTRIBUTION OF SMALL PRIME GAPS

BY

R. P. BRENT

TECHNICAL REPORT NO. CS 123
FEBRUARY 28, 1969

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

Empirical Evidence for a Proposed Distribution of Small Prime Gaps

By

R. P. Brent

Computer Science Department

Stanford University*

---

*Most of the work described here was done at Monash University, Australia. The reproduction was sponsored by the Office of Naval Research and the National Science Foundation. The author is at present supported by a CSIRO fellowship.

Introduction

There are many unsolved problems concerning the distribution of prime numbers. For example, it is not known whether there is an infinity of 'twins', pairs p and $p + 2$ both prime, although empirical evidence strongly suggests that there is (see [1]). In this paper the broader question of the distribution of small even gaps between successive large primes is investigated. The arguments used involve statistical assumptions which, although intuitively reasonable, are not, and perhaps can not be, rigorously justified. Hence the results obtained are not formally proven. They are, however, very well supported by extensive empirical evidence. Hence e merit claimed for the results of this paper is that, theoretically justifiable or not, they give an extremely good representation of the actual distribution of small prime gaps. Considering the irregularities of this distribution (see Diagram 1), any reasonable explanation of it is interesting.

Notation

This section is probably best referred to when needed below.

Throughout let Q be the set of odd primes $3,5,7,11,\ldots$, and let $q \in Q$. Let N be a large integer; p, a (varying) prime with $p \simeq N$; and r, a small positive integer.

V is the set of all r-tuples $v = (v_1,\ldots,v_r)$, where each v_i is 0 or 1 and $v_r = 1$.

For $k \geq 1$, define

$$c_k = \prod_{q>k+1} \left(\frac{1-1/(q-k)}{1-1/q}\right) = \prod_{q>k+1} \left(1 - \frac{k}{(q-1)(q-k)}\right)$$

and, for $r \geq k$, define

$$F_{r,k} = \frac{-\left(-2 \prod_{q\leq r+1} (q/(q-1))\right)^k c_1 c_2 \cdots c_k}{\prod_{q\leq r+1} \prod_{i=1}^{\min(k,q-2)} (1-i/((q-1)(q-i)))}$$

For $v \in V$, let the nonzero components of v be, in order, $v_{n_1}, v_{n_2}, \ldots, v_{n_k}$ (so $n_k = r$), and let $n_0 = 0$.

If L is the set of $n_j \pmod q$ for $j = 0, 1,\ldots, i-1$ then define

$$g(q,i,v) = \begin{cases} 0 & \text{if } n_i \pmod q \in L \\ \dfrac{1}{q-|L|} & \text{otherwise.} \end{cases}$$

Finally, let

$$h(v) = \prod_{q \le r+1} \prod_{i=1}^{k} (1-g(q,i,v))$$

and

$$S_{r,k} = \sum_{v \in V,\ \sum_1^r v_i = k} h(v) .$$

The notation $m \nmid n$ means that n is not divisible by m .

Theory

Everywhere "the probability of event E given F", written $P(E|F)$, should be interpreted as relative frequency, in a sense which should be clear from the context.

We are concerned with finding a function $f(r)$ which approximates the probability that a prime gap in a given region will be of length $2r$. More precisely, if M is an integer, large compared to r and $\log N$, but small compared to N, and if there are $n + 1$ primes in the interval $(N-M, N+M)$, and if m of the gaps between consecutive primes in this interval are of length exactly $2r$, then we expect that

$$m/n \doteqdot f(r) .$$

The point of this paper is the substantiation of:

Conjecture 1

Let $A_{r,k} = F_{r,k} \cdot S_{r,k}$, where $F_{r,k}$ and $S_{r,k}$ are defined above. Then for small r, i.e. $r \lessdot \log N$, a function f satisfying the conditions of the previous paragraph is

$$f(r) = \sum_{k=1}^{r} \frac{A_{r,k}}{(\log N)^k} .$$

(Table 1 gives some computed values for the $A_{r,k}$.)

Before discussion the Conjecture, it is interesting to deduce some of its immediate consequences:

Corollary 1

For fixed r ,

$$f(r) = \frac{A_{r,1}}{\log N} \cdot (1+o(1)) \quad \text{as} \quad N \to \infty .$$

The proof is immediate. Note that, from the definition of $A_{r,k}$, we have

$$A_{r,1} = 2c_1 \prod_{q|r} \left(\frac{q-1}{q-2}\right) ,$$

and as $\Pi(\frac{q-1}{q-2})$ diverges the $A_{r,1}$ are unbounded.

In the following, by $a \sim b$ we always mean that

$$\lim_{N \to \infty} a/b = 1 .$$

Corollary 2

If $\eta(r)$ is the number of pairs of consecutive primes p and $p + 2r$ with $p < N$, then

$$\eta(r) \sim \frac{A_{r,1} \cdot N}{(\log N)^2}$$

Proof

From Corollary 1 and the prime number theorem, we see that

$$\eta(r) \sim \int_c^N \left(\frac{A_{r,1}}{\log t}\right) \frac{dt}{\log t}$$

and integration by parts gives the result.

Corollary 3

Putting 1 for r in Corollary 2, the number of twin primes less than N is

$$\sim \frac{2c_1 \cdot N}{(\log N)^2} .$$

Again I would emphasize that Corollaries 1-3, while following rigorously from Conjecture 1, have not been proven, for they depend on the informal arguments used below to substantiate (not prove) Conjecture 1.

Before discussing Conjecture 1, we need some definitions and a Lemma. Let $v \epsilon V$, and p range over the primes near N as before. For $r' \leq r$, define

$$q(r',v) = P(1 \leq i \leq r' \wedge v_i = 1 \supset p+2i \epsilon Q)$$

and

$$\bar{q}(r',v) = P(1 \leq i \leq r' \supset (p+2i \epsilon Q \equiv v_i = 1)) ,$$

where parentheses may be restored by the usual conventions.

We shall abbreviate $q(r,v)$ by $q(v)$ and $\bar{q}(r,v)$ by $\bar{q}(v)$. Define

$$s(v) = (-1)^{\sum_1^{r-1} v_i} .$$

If v, $v' \epsilon V$ we write $v' \geq v$ if $v_i' \geq v_i$ for each $i = 1, \ldots, r$.

We shall see below that it is possible to estimate $q(v)$, so we need to express the function f in terms of the $q(v)$. The following Lemma does this:

Lemma

$$f(r) \doteq \sum_{v \in V} s(v) \cdot q(v) .$$

Proof

From the definition of $\bar{q}$ we have

$$f(r) \doteq \bar{q}((0,0,\ldots,0,1)) , \qquad (1)$$

but from the definition of q it is easy to see that

$$q(v) = \sum_{v' \geq v} \bar{q}(v') .$$

Hence

$$\sum_{v \in V} s(v) q(v) = \sum_{v' \in V} \left(\bar{q}(v') \cdot \sum_{v \leq v'} s(v) \right) . \qquad (2)$$

But

$$\sum_{v \leq v'} s(v) = \binom{k'-1}{0} - \binom{k'-1}{1} + \ldots + (-1)^{k'-1} \binom{k'-1}{k'-1}$$

$$= \begin{cases} 0 & \text{if } k' \neq 1 \\ 1 & \text{if } k' = 1 \end{cases} .$$

Hence the result follows from (1) and (2).

Now we are ready to complete the substantiation of Conjecture 1. From the definition of conditional probability, we see that

$$\frac{q(r,v)}{q(n_{k-1},v)} = P(p+2r\in Q \mid 1\le i<k \supset p+2n_i \in Q)$$

$$= P(q\in Q \wedge q<p \supset q \nmid p+2r \mid 1\le i<k \supset p+2n_i \in Q) \ .$$

At this stage we make an assumption which, although reasonable, is really only justified by the agreement of Conjecture 1 with empirical data. We assume independence of divisibility by the different primes q in the above expression. Actually, it is enough to assume that this is a good approximation for primes q small compared to p. The assumption gives

$$\frac{q(r,v)}{q(n_{k-1},v)} \doteq \prod_{q<p} P_q \ , \qquad (3^*)$$

where

$$P_q = P(q \nmid p+2r \mid 1\le i<k \supset p+2n_i \in Q) \ . \qquad (4)$$

We now make a rather similar assumption, that the condition $p+2n_i \in Q$ only affects P_q in that it assures that $q \nmid p+2n_i$. This gives

$$P_q = P(q \nmid p+2r \mid 1\le i<k \supset q \nmid p+2n_i) \ , \qquad (*)$$

$$= 1 - P(q \mid p+2r \mid 1\le i<k \supset q \nmid p+2n_i) \ ,$$

but considering the possibilities for $p+2r \pmod q$, bearing in mind that p, being prime, is not divisible by q, and looking back to the definition of g, it is not difficult to see that the last term is just $g(q,k,v)$. Hence

$$P_q = 1 - g(q,k,v) \ . \tag{5}$$

Since p is odd, the prime number theorem gives

$$\frac{2}{\log N} \doteqdot P(p+2r\epsilon Q)$$

$$= P(q\epsilon Q \wedge q<p \supset q \nmid p+2r) \ .$$

By another assumption similar to those above this is

$$\prod_{q<p} P(q \nmid p+2r) = \prod_{q<p} (1-1/q) \ , \tag{*}$$

so

$$\prod_{q<p} (1-1/q) \doteqdot \frac{2}{\log N} \tag{6}$$

Combining (3) to (6) gives

$$\frac{q(r,v)}{q(n_{k-1},v)} \doteqdot \frac{2}{\log N} \prod_{q<p} \frac{1-g(q,k,v)}{1-1/q} \ . \tag{7}$$

Observe that if $q > r$ then

$$g(q,k,v) = 1/(q-k) ,$$

and if $q > r + 1$ then, since $k \leq r$, this is < 1. Now the product

$$\prod_{q>k+1} \left(\frac{1-1/(q-k)}{1-1/q}\right)$$

converges, and we assumed that $p \sim N$ was large, so in (7) the condition $q < p$ may be dropped. Also, since $q(0,v) = 1$, we have

$$q(r,v) = \frac{q(r,v)}{q(n_{k-1},v)} \cdot \ldots \cdot \frac{q(n_1,v)}{q(0,v)} ,$$

so from (7)

$$q(r,v) \doteq \left(\frac{2}{\log N}\right)^k \prod_{i=1}^{k} \prod_{q \in Q} \left(\frac{1-g(q,i,v)}{1-1/q}\right) \qquad (8)$$

Now substitution of (8) into the result of the Lemma, and a rearrangement of the products using the observation about g above, gives the required result. Steps where statistical assumptions were made are indicated by (*).

Empirical Tests

First it was necessary to evaluate the constants $A_{r,k}$. The c_k for $k = 1, 2, \ldots, 40$ were calculated by taking the product over primes less than 40000, and roughly approximating the remainder by an integral. The first few are $c_1 = 0.66016$, $c_2 = 0.72160$, $c_3 = 0.48412$, $c_4 = 0.65085$, $c_5 = 0.45529$, $c_6 = 0.71314$, $c_7 = 0.62911$, $c_8 = 0.51704$, and $c_9 = 0.34787$. Computation of the $A_{r,k}$ is more interesting. Difficulties soon arise because of the large number of terms in the sum $S_{r,k}$ when k is large (in fact when k is not very small). The $A_{r,k}$ were computed by a straightforward method for $r \le 18$, $k \le r$, and also for $r = 19$, 20, 21, $k \le 8$. See Table 1. An interesting combinatorial problem, which we shall not discuss here, is the computation of the function $u(r) = \max\{k \le r \mid A_{r,k} \ne 0\}$.

Eleven blocks, each of about 8.10^6 numbers and in the region from 6.10^6 to 2.10^{10}, were searched for primes, and for each block the actual distribution of gaps was found. Taking for N the midpoint of the block (this is not critical), the probabilities $f(1)$, $f(2), \ldots, f(21)$ and $1 - \sum_1^{21} f(r)$ were calculated from the $A_{r,k}$ and Conjecture 1. The 'predicted distribution' was just these probabilities multiplied by the total observed number of gaps (so one degree of freedom is lost), and the predicted and actual distributions were compared. In no case did the χ^2 test indicate a significant difference at the 5% (or even at the 10%) level. Generally, the fit seemed slightly better than chance, which is perhaps reasonable on intuitive grounds, but in only three of the eleven cases was χ^2_{21} significantly <u>small</u>

at the 5% level. The intervals, number of primes in them, χ^2_{21} for 21 degrees of freedom, and probability of such χ^2_{21} being exceeded in sampling from identical populations, are shown in Table 2.

The method of searching for primes was a sieve method similar to that described in [3]. Primes up to the square root of the largest number to be tested are first found by some method, and then blocks of numbers are 'sieved'. Only odd numbers are considered, and only a one bit flag for each number is necessary. Actually it is quicker to use the smallest addressable unit. The blocks should be as large as possible. On a CDC 3200 with 15 bit index registers (with sign extension to 17 bits for character addressing) and 1's complement arithmetic, a block length of $2^{15}-1$ can be used, and the innermost loop is only three instructions with one storage reference. The method is very fast compared, say, to the ALGOL procedures [2]. Around 10^7 the time to search a million numbers and output the roughly 60,000 primes to tape (for possible future use) was 20.1 seconds, around 10^{10} this increased to 30.4 seconds. The program was checked using the amazingly accurate tables [4], and all computing was done on a CDC 3200 at Monash University.

In a typical case, 347570 primes were found (in 243 sec.) in the interval $(10^{10}, 10^{10}+8{,}000{,}074)$. The distribution of gaps is shown in Diagram 1, and Table 3 compares the actual and predicted distributions. Note the approximate equality of the peaks for gaps 2 and 4, the high peak for 6, and the general irregularity of the distribution, which are typical of all eleven cases, and as predicted by Conjecture 1.

Conclusion

Using Conjecture 1 and the constants $A_{r,k}$ in Table 1, the distribution of small prime gaps predicted was in good agreement with empirical results for over 4,000,000 gaps. As the distribution is so irregular, which can be seen by a glance at Diagram 1, it is hard to believe that this good fit is just a coincidence. Hence any results to be proved concerning, say, twin primes, will probably have to be compatible with Conjecture 1, or at least with Corollary 3.

Table 1

r	k=1	2	3	4	5	6	7	8	9	10	11
r=1	1.3203										
2	1.3203	0									
3	2.6406	-5.7165	0								
4	1.3203	-5.7165	4.1512	0							
5	1.7604	-8.5747	8.3023	0	0						
6	2.6406	-20.008	41.512	-20.264	0	0					
7	1.5844	-14.291	38.744	-30.395	0	0	0				
8	1.3203	-14.291	49.814	-60.790	17.298	0	0	0			
9	2.6406	-32.870	138.37	-222.90	103.79	0	0	0			
10	1.7604	-27.868	160.51	-405.27	415.16	-107.94	0	0			
11	1.4670	-22.509	124.93	-295.51	249.10	0	0	0			
12	2.6406	-48.590	343.56	-1161.8	1868.2	-1133.4	0	0			
13	1.4404	-29.869	243.58	-989.75	2087.7	-2140.9	831.88	0			
14	1.5844	-33.048	270.42	-1097.6	2290.3	-2266.8	792.27	0			
15	3.5209	-86.248	855.67	-4408.2	12542.	-19272.	14320.	-3780.3			
16	1.3203	-36.300	413.65	-2532.1	9022.4	-18953.	22649.	-13945.	0	0	0
17	1.4083	-39.046	448.96	-2771.5	9927.3	-20794.	24340.	-14176.	3409.4	0	0
18	2.6406	-79.332	1000.5	-6889.0	28204.	-70154.	104110	-87178.	3117.2	0	0
19	1.3980	-44.642	600.71	-4425.9	19396.	-51300.	78886.	-63103.	36919.	-6124.0	0
20	1.7604	-58.135	815.12	-6321.4	29583.	-85427.	149070	-146590			
21	3.1688	-115.20	1803.0	-15890.	86545.	-300640	662380	-888200			
(22	1.4670	-56.513	946.23	-9036.7	54285.	-213220)					

The constants $A_{r,k}$. The last digit may be in error by 2 or 3, especially for higher k. Values which are omitted are zero for $r \leq 18$.

Table 2

$\log_{10} N$	a	b	n + 1	χ^2_{21}	$P(\chi^2 \geq \chi^2_{21})$
7.00	6.10^6	8,000,034	497230	15.28	0.81
7.38	2.10^7	8,000,098	470830	14.08	0.87
7.81	6.10^7	8,000,040	445230	15.55	0.79
8.31	2.10^8	8,000,022	418280	18.79	0.60
8.78	6.10^8	8,000,078	395930	8.73	0.991
9.00	1.10^9	8,000,198	386000	21.69	0.42
9.30	2.10^9	8,000,000	374240	27.03	0.17
9.78	6.10^9	8,000,004	355150	9.20	0.987
10.00	1.10^{10}	8,000,074	347570	15.54	0.79
10.18	15.10^9	8,000,000	341390	19.36	0.56
10.30	2.10^{10}	8,000,000	337310	10.99	0.96

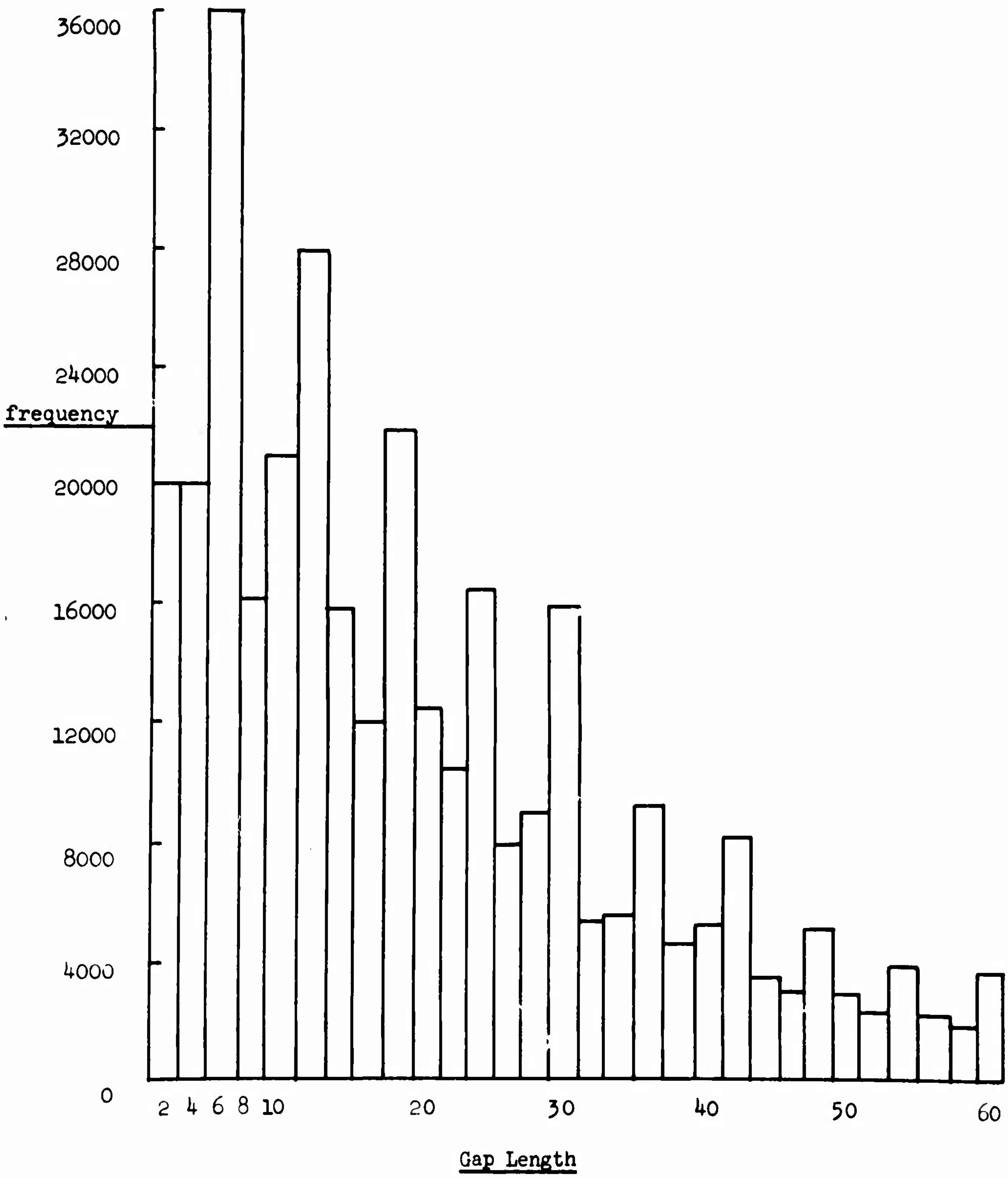
Empirical results for distribution of prime gaps. The interval searched is (a, a+b) with midpoint N, number of primes in interval is n+1 (so n gaps). Testing the fit of actual and predicted distribution of gaps of length 2, 4, ..., 42 and remainder gives χ^2_{21} with 21 degrees of freedom.

Table 3

r	f_o	f_e	$(f_o-f_e)/\sqrt{f_e}$
1	19943	19930	+0.09
2	19977	19930	+0.34
3	36145	36112	+0.17
4	16325	16300	+0.19
5	21054	21188	-0.92
6	28009	27900	+0.65
7	15783	15613	+1.36
8	11973	11905	+0.62
9	21956	21981	-0.18
10	12403	12395	+0.07
11	10510	10593	-0.81
12	16435	16449	-0.11
13	7810	7979	+0.34
14	8896	8710	+1.99
15	15957	16147	-1.50
16	5249	5222	+0.38
17	5533	5504	+0.38
18	9200	9183	+0.18
19	4428	4397	+0.47
20	5215	5257	-0.58
21	8033	8007	+0.29
22,...,130	46735	46867	-0.61

Distribution of the 347,569 prime gaps in the interval (10^{10}, 10,008,000,074). For a gap of length 2r the actual frequency is f_o and the predicted frequency f_e (with equal totals). The X^2 test gives P = 0.79, so does not show a significant difference between the two distributions.

Diagram 1



The frequency of occurrence of small prime gaps in the interval $(10{,}000{,}000{,}000,\ 10{,}008{,}000{,}074)$.

References

[1] Hardy, G. H., & Wright, E. M. An Introduction to the Theory of Numbers, Oxford University Press, Oxford, 1962.

[2] Chartres, B. A. Algorithms 310, 311. Comm. ACM 10, 9 (Sept. 1967), 569-570.

[3] Howland, J. E. Letter to the Editor. Comm. ACM 11, 3 (March 1968), 149.

[4] Lehmer, D. H. List of Prime Numbers from 1 to 10,006,721, Washington, 1914.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Computer Science Department Stanford University Stanford, California 94305	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP ---
3. REPORT TITLE EMPIRICAL EVIDENCE FOR A PROPOSED DISTRIBUTION OF SMALL PRIME GAPS	
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Manuscript for Publication (Technical Report)	
5. AUTHOR(S) (First name, middle initial, last name) Richard P. Brent	
6. REPORT DATE February 28, 1968	7a. TOTAL NO. OF PAGES: 18 — 7b. NO. OF REFS: 4
8a. CONTRACT OR GRANT NO. N00014-67-A-0112-0029 b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) CS 123 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) none
10. DISTRIBUTION STATEMENT Releasable without limitations on dissemination.	
11. SUPPLEMENTARY NOTES ---	12. SPONSORING MILITARY ACTIVITY Office of Naval Research

13. ABSTRACT

The distribution of small gaps between adjacent prime numbers is studied. A model for this distribution is derived from probability arguments. Empirical evidence strongly supports this model. An asymptotic density function for twin primes is suggested, and support is given to the conjecture that there is an infinity of twin primes.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6801

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
prime numbers						
number theory						
twin primes						
prime gaps						
distribution						
sieving						
Eratosthenes						
combinatorial						